Asian Institute of Technology
School of Engineering and Technology

AT84.02 : Business Intelligence and Analysis

Instructor:
Dr.Vatcharaporn Esichaikul

Project Report on
Business Intelligence System for Music Streaming Platform

Group Members
Sunny Kumar Tuladhar (st122336)
Min Khant Soe (st122277)
Amanda Shrestha (st122245)
Win Win Phyo (st122314)

Date of Submission:
23rd April, 2022

Table of contents

Introduction

The main goal of this project is to create a business intelligence system for a music streaming platform to be able to provide insights for the streaming platform and recommendations for the artists and the listeners.

Streaming platform is basically an application which hosts the songs of many different musical artists which the listeners can listen to after paying a subscription fee. Listeners can listen to any artist they want by searching for that particular artist. When the listener does not actively search for the songs they want to listen to, the platform will recommend a song they like based on their listening preferences and other metrics such as trending genres and trending songs.

Objectives

The main goal of the system is to increase the usage of the application i.e. to get the listener to listen to songs more often using this application , increase the number of users and then generate more revenue for the platform. We achieve this goal by using our BIS model such that the audience gets recommended the best songs that they might like.

We achieve the goal by doing the following tasks

1. To visualize useful data and current trends from artists , songs and users
2. To analyze data from visualization
3. Create a prediction model to predict whether a user will like a song

We will be creating visualizations and dashboards from the available data , analyzing them and making decisions such that we can maximize listening time for the audience. We also create a 'Song Likeness' Prediction model that will be able to predict whether a user will like the given song or not.

End Users of the System

We separated the system's end users into three main categories. We have targeted three main end users who will be given insights of the platform in the form of a dashboard in order to maximize listening time for the audience.

1. **Audience**: who listen to the songs and purchase subscriptions to the streaming platform. These are the main source of revenue for the company.
2. **Artists**: These are the creators of the content of the platform. Their music is what the audience will be listening to. The company will pay them according to their number of listens.
3. **Company**: This is the main company to which the platform belongs to. All the revenue goes to this company.

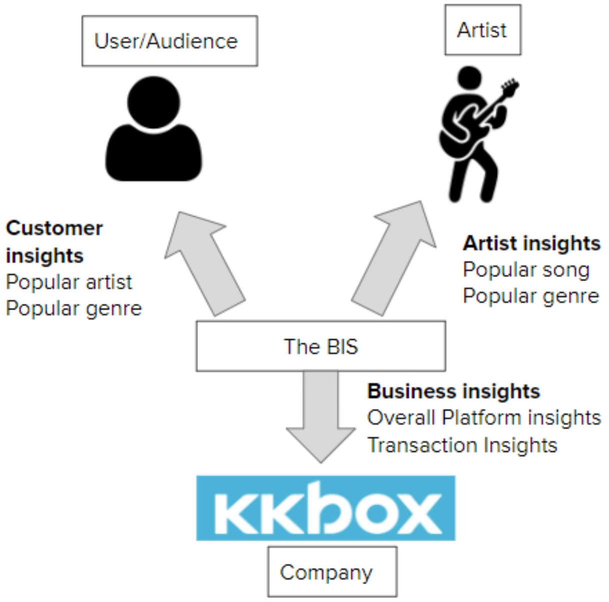


Figure1 : Target users of the system

Software for implementation

Tableau

Tableau is a data visualization tool that is commonly used for Business Intelligence, but it can also be used for other purposes. It assists in the creation of interactive graphs and charts in the form of dashboards and spreadsheets in order to get business insights.

We choose Tableau for its advantages and usefulness. Firstly, we can create an interactive visual representation in a short period of time by using the drag and functions. Tableau is more comfortable to be used than Power BI; we do not need specific technical skills to implement and visualize the data on Tableau. Thus, the organizations can reduce the numbers of employees for the data visualization and save the expense. In other words, Tableau removes the need of very professional IT staff to visualize the data. Tableau organizes and analyzes data in a logical and easy-to-understand manner. This aids in the reduction of development time and the speeding up of decision-making. The capacity to adjust to changing market conditions is a significant competitive advantage. Moreover, Tableau assists the organization in making quick decisions based on the data, resulting in a competitive advantage. With a single click, Tableau's built-in "Show-Me" function changes between a range of chart formats. There's no need to modify data or spend hours preparing and aligning elements for each chart style. Next, we can manage large amounts of data and create different types of visualization without disturbing the performance of the dashboards. Tableau doesn't have a row limit, so the organizations can import and analyze millions of rows of data. It's also scale-ready, so it removes the concern about excessively long processing times. All of this provides the firms with a more comprehensive view of the business within a short period of time. Besides, Tableau can be used to connect with different data sources like SQL, etc. and also be used to work with Python or R to do complex calculations, remove the load of the software by performing data cleansing tasks and avoid performance issues. Firms can also run Tableau and visualize data on mobile apps for either IOS or Android. Additionally, it also has strong mobile support. Furthermore, Tableau has numerous resources, guides, trains, and forums on the Internet, so it is easy to get information when the firms face some problems using it. Lastly, Tableau's price is reasonable and affordable, so it reduces the need of large expense on data visualization.

Python

We use python for data preprocessing and to create and train the Machine Learning algorithm for the project. Python and its immense amount of libraries for machine learning make it easier to preprocess and train these models with ease. The csv files were processed using the pandas library. Python can handle huge amounts of data with ease compared to conventional software. The data we used in this project amounted to around 1.5GB in csv files which is quite big. Using the pandas library we manipulated this large dataset into a numpy array. Tables were merged, categorical values converted to numerical and then finally converted to numpy. This numpy array was then fed into machine learning models. These models were created using the sci-kit learn and the xgboost library.

Dataset Information

In this project, we collected two datasets of KKBOX which are published in Kaggle. These two datasets are *WSDM - KKBox's Music Recommendation* Challenge and *WSDM - KKBox's Churn Prediction Challenge*. From these two datasets, we used the following data;

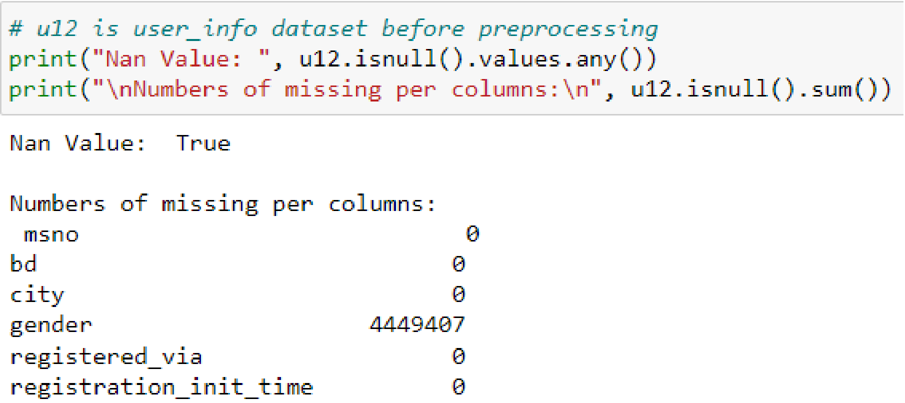
- **User Information** which include user id, age, gender, city, registration method, initial registered time.
- **Transactions Information** which include user id, payment method, plan of the subscription, price of the plans, actual amount paid, auto_renew and manual buy of subscription, cancellation of subscription and transaction date.
- **User and Song Relation Information** which include user id, song id, source_system_tab which is the name of the tab where the event was triggered, source_screen_name which is the name of the layout a user sees, source_type which is an entry point a user first plays music on mobile apps, and the data whether the song is listened more than one time.
- **Song Information** which include song id, length of the song, genre category, name of artist, name of composer, name of lyricist and language of the song
- **Song Extra Information** which include song id, song name, and international recording code.

Data Preprocessing

We used python programming for data engineering. First, we combined the user information of these two datasets and checked whether the user information of these two datasets are duplicated; if user ids are duplicated, we remove one. The reason for combining the data is because we do not want any missing information between these two datasets. We checked the null values from all the data and filled them with relevant data.

User Information

For user information, we found that gender is filled with null values.

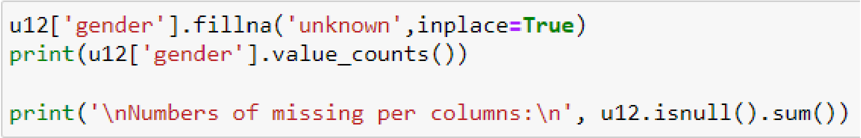


```
# u12 is user_info dataset before preprocessing
print("Nan Value: ", u12.isnull().values.any())
print("\nNumbers of missing per columns:\n", u12.isnull().sum())
```

```
Nan Value:  True

Numbers of missing per columns:
 msno                            0
bd                              0
city                            0
gender                    4449407
registered_via                  0
registration_init_time          0
```

Figure 2.1: Checking the Null values for user information

So, we filled those null values with ‘unknown” insteading of manipulating data with specific values such as “male” and “female”.

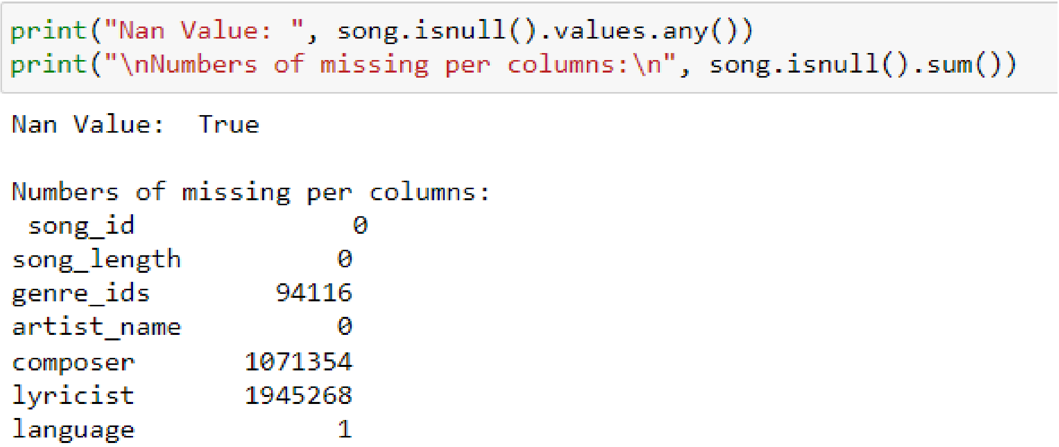


```
u12['gender'].fillna('unknown',inplace=True)
print(u12['gender'].value_counts())

print('\nNumbers of missing per columns:\n', u12.isnull().sum())
```

Figure 2.2 : Filling the Null values with relevant data in user information

Song Information

For song information, we found that genre_ids, composer, lyricist and language are filled with null values.

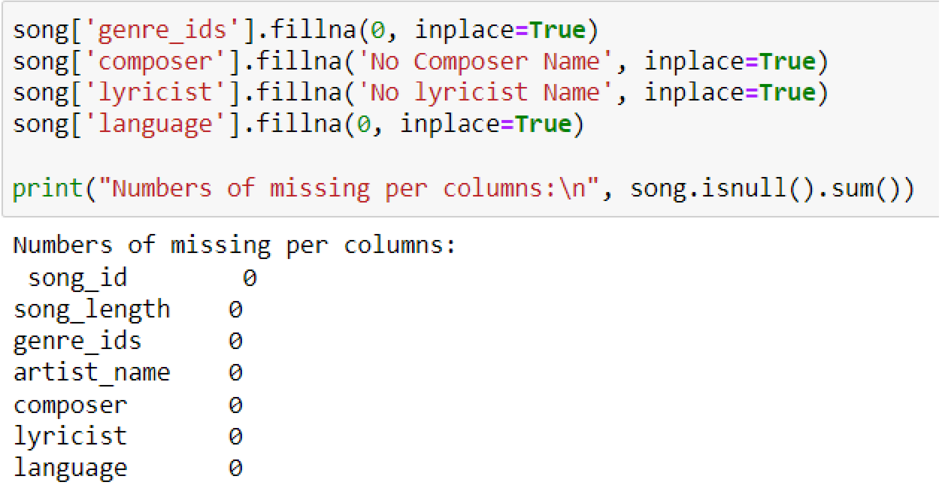


```
print("Nan Value: ", song.isnull().values.any())
print("\nNumbers of missing per columns:\n", song.isnull().sum())
```

```
Nan Value:  True

Numbers of missing per columns:
 song_id               0
song_length           0
genre_ids         94116
artist_name           0
composer        1071354
lyricist        1945268
language              1
```

Figure 3.1 : Checking the Null values for song information

We filled those null values for genre_ids, composer,lyricist and language with “0”, “No Composer Name”, “No lyricist Name” and “0” respectively.



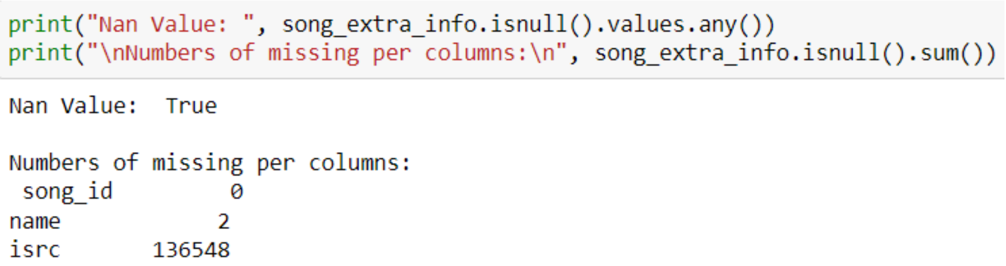
```
song['genre_ids'].fillna(0, inplace=True)
song['composer'].fillna('No Composer Name', inplace=True)
song['lyricist'].fillna('No lyricist Name', inplace=True)
song['language'].fillna(0, inplace=True)

print("Numbers of missing per columns:\n", song.isnull().sum())
```

```
Numbers of missing per columns:
 song_id        0
song_length    0
genre_ids      0
artist_name    0
composer       0
lyricist       0
language       0
```

Figure 3.2 : Filling the Null values with relevant data in song information

Song Extra Information

For song extra information, we also found that name and isrc are filled with null values.

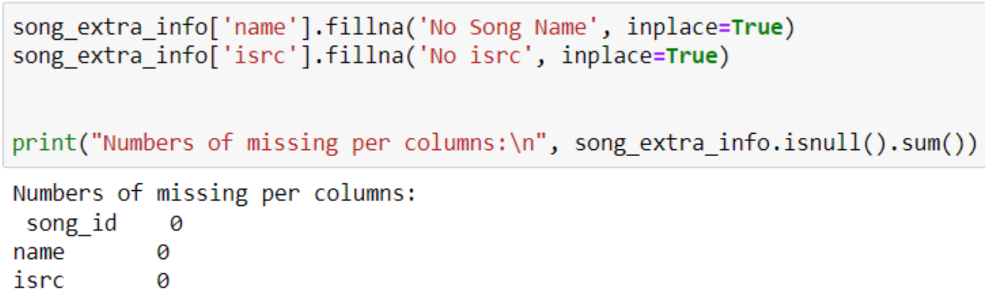


```
print("Nan Value: ", song_extra_info.isnull().values.any())
print("\nNumbers of missing per columns:\n", song_extra_info.isnull().sum())
```

```
Nan Value:  True

Numbers of missing per columns:
 song_id         0
name            2
isrc       136548
```

Figure 4.1 : Checking the Null values for song extra information

We filled those null values for name and isrc with “No Song Name” and “No isrc” respectively.



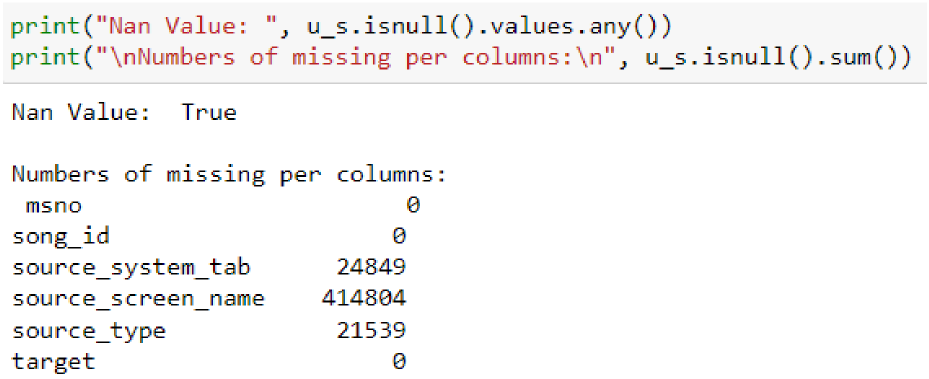
```
song_extra_info['name'].fillna('No Song Name', inplace=True)
song_extra_info['isrc'].fillna('No isrc', inplace=True)


print("Numbers of missing per columns:\n", song_extra_info.isnull().sum())
```

```
Numbers of missing per columns:
 song_id    0
name       0
isrc       0
```

Figure 4.2 : Filling the Null values with relevant data in song extra information

User and song relation information

For user and song relation information, we found that source_system_tab, source_screen_name and source_type are filled with null values.

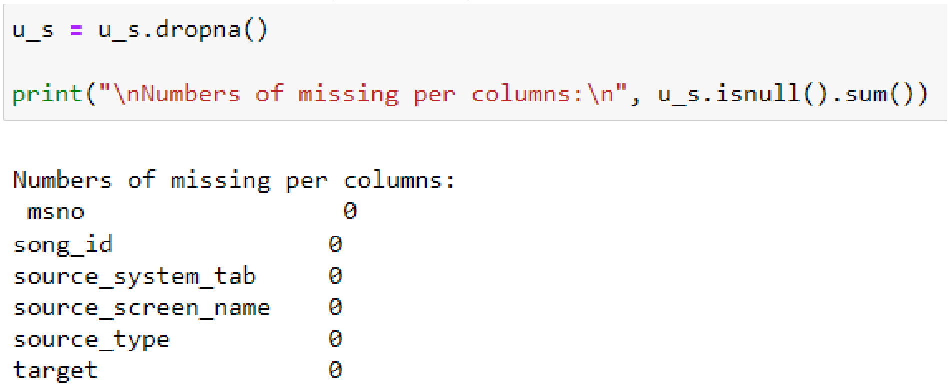


```
print("Nan Value: ", u_s.isnull().values.any())
print("\nNumbers of missing per columns:\n", u_s.isnull().sum())
```

```
Nan Value:  True

Numbers of missing per columns:
 msno                       0
song_id                    0
source_system_tab      24849
source_screen_name    414804
source_type            21539
target                     0
```

Figure 5.1 : Checking the Null values for user and song relation information

We removed these null values data for simplicity of the data for prediction with machine learning and there is no effect to other tables by removing them.



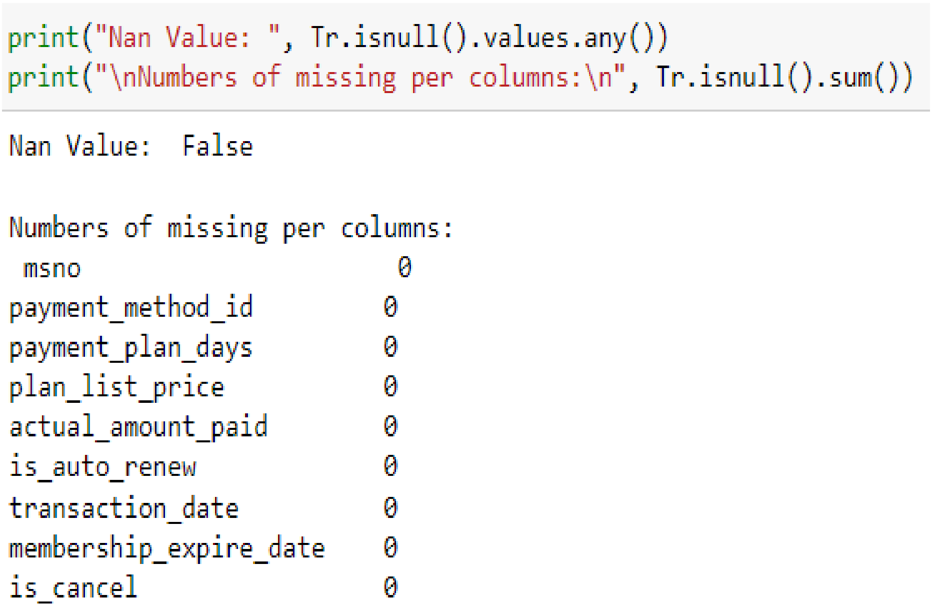
```
u_s = u_s.dropna()

print("\nNumbers of missing per columns:\n", u_s.isnull().sum())
```

```
Numbers of missing per columns:
 msno                 0
song_id              0
source_system_tab    0
source_screen_name   0
source_type          0
target               0
```

Figure 5.2 : Filling the Null values with relevant data in user and song relation information

Transaction Information

For transaction information, there is no missing data.



```
print("Nan Value: ", Tr.isnull().values.any())
print("\nNumbers of missing per columns:\n", Tr.isnull().sum())
```

```
Nan Value:  False

Numbers of missing per columns:
 msno                     0
payment_method_id        0
payment_plan_days        0
plan_list_price          0
actual_amount_paid       0
is_auto_renew            0
transaction_date         0
membership_expire_date   0
is_cancel                0
```

Figure 6.1 : Checking the Null values for transaction information

After that, to visualize the data, we loaded the data file into Tableau and linked the relationship of tables.

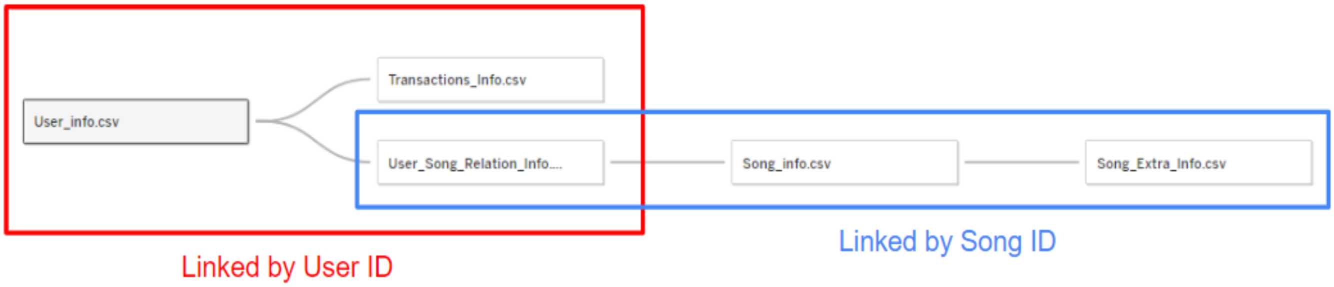
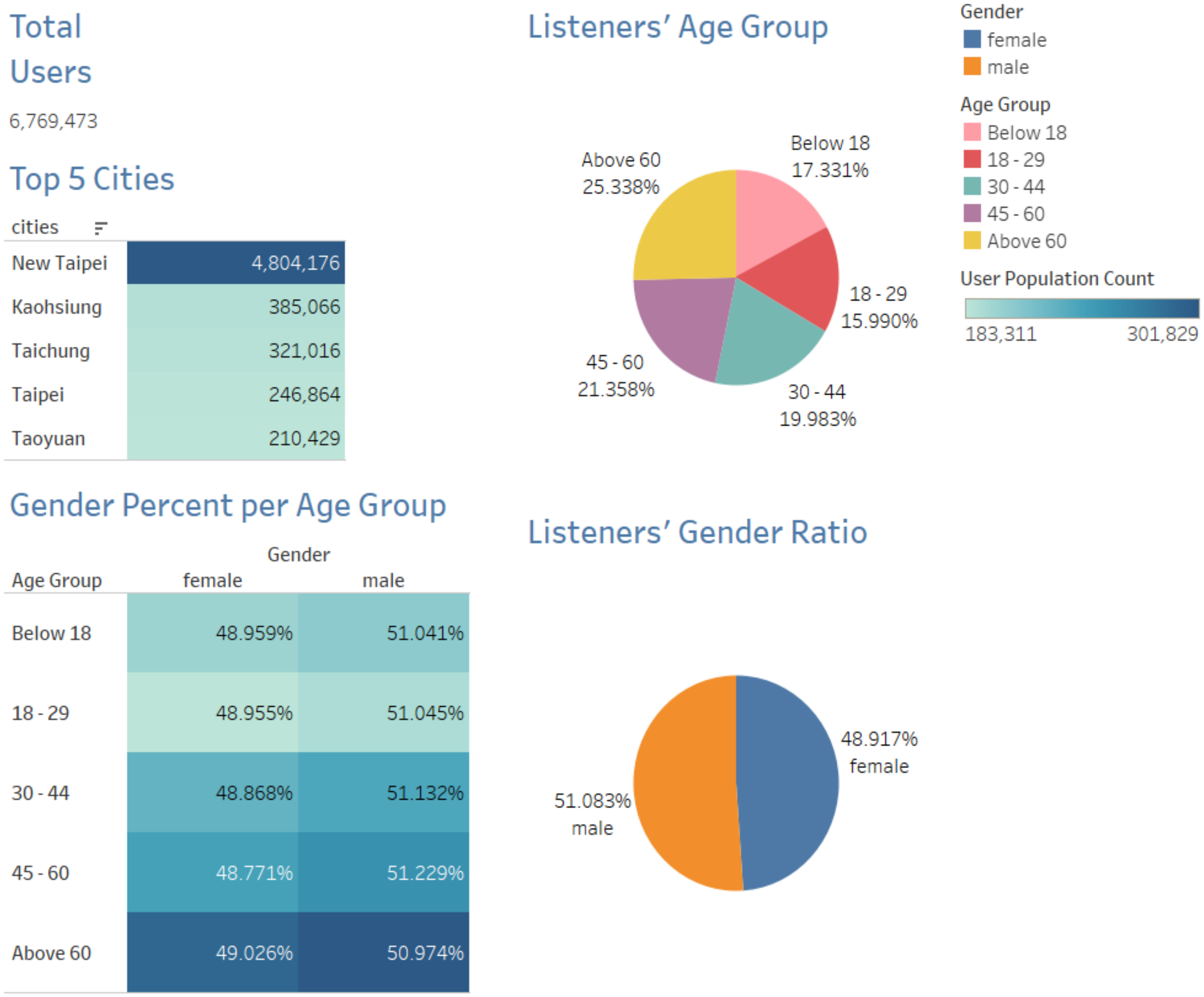


Figure 7 : Relationship between the data tables

Since these data are not for machine learning tasks, the names of cities, genre category and language of songs are presented in number format. so we have to manually fill text representation of them which were found by researching on the internet.

Descriptive Analysis

This part of the analysis deals with recognising the current and past situation of the platform. Here we observe and analyze various aspects of the audience demographic, their listening habits and find methods to improve the overall engagement in the platform by giving insights to the three end users. These insights are presented in the form dashboard visualizations as shown below.



Total Users

6,769,473

Top 5 Cities

cities	
New Taipei	4,804,176
Kaohsiung	385,066
Taichung	321,016
Taipei	246,864
Taoyuan	210,429

Gender Percent per Age Group

	Gender	
Age Group	female	male
Below 18	48.959%	51.041%
18 - 29	48.955%	51.045%
30 - 44	48.868%	51.132%
45 - 60	48.771%	51.229%
Above 60	49.026%	50.974%

Figure 8 : Customer Demographic Dashboard

The dashboard in Figure 8 illustrates the different demography of the users of the music streaming platform. This dashboard helps the music company understand its users, and to create better strategies to boost the users' satisfaction and experience. According to this dashboard, the total number of its customers is more than 6.7 millions and the majority of its customers are from New Taipei. By analyzing the major cities of the company's users in Taiwan, the company's music platform is most popular in New Taipei, Kaohsiung, Taichung, Taipei and Taoyuan.

Interestingly, most of its users are people with an age of above 60 which is the retired age whereas the age between 18 to 29 is found to be the lowest listener age in this music platform. According to the article, *Education in Taiwan* by Jessica Magaziner, Research Associate in 2016, WES, the people of Taiwan strongly focus on the education of their children. The age between 18 to 29 can be considered as university student age. Hence, they will not have much time to spend time on the music and money to buy the subscription of KKBOX. In addition, the age between 45 to 60 is the second largest and followed by the age between 30 to 44; both of them can also be considered as the middle age. When we combine these two groups as one, the middle age group becomes 41.34 percent which is much more than the percent of retired age. From this visualization, we can analyze that this KKBOX music streaming platform is famous for middle-aged people. Thus, from the analysis of the listeners' age group visualization and the Education in Taiwan article, we have a final assumption; since people of Taiwan care a lot about the education and since the middle-aged people are the highest income earner, it is possible that they buy the subscription of this music platform for their children and limit their children's spend time on entertainment including music and as a result, the middle age group becomes the largest group of user in this platform.

Next, The difference in ratio between male and female users is 21.66 percent which can be interpreted as the gender diversity of the users is insufficient. Moreover, with the accordance of gender percent per age group data visualization, male percent is slightly higher in all age groups. Thus, the data can be interpreted as this music streaming platform is more popular for male users in Taiwan.

This dashboard can also be used for advertisement purposes on the music streaming platform. By providing this demographic data to the public, the music streaming company can attract other businesses to advertise on its music streaming platform; the music streaming company can find sponsorship from those advertisements and it will increase the company's revenue and profits. The sponsor companies can find their target audience by observing this dashboard which includes total numbers of users, the top cities where the users live, and the age group and gender ratio of the users. Furthermore, the company can also earn money by selling membership subscriptions which will not be affected by any advertisement while listening to the music. The advertisement of the other company who chose to present their product and service on this music streaming platform will pop up to the users who are using it for free; by doing this, the music company can deliver the brand message to the target customers of those sponsor companies; it will also urge the users to buy the membership.

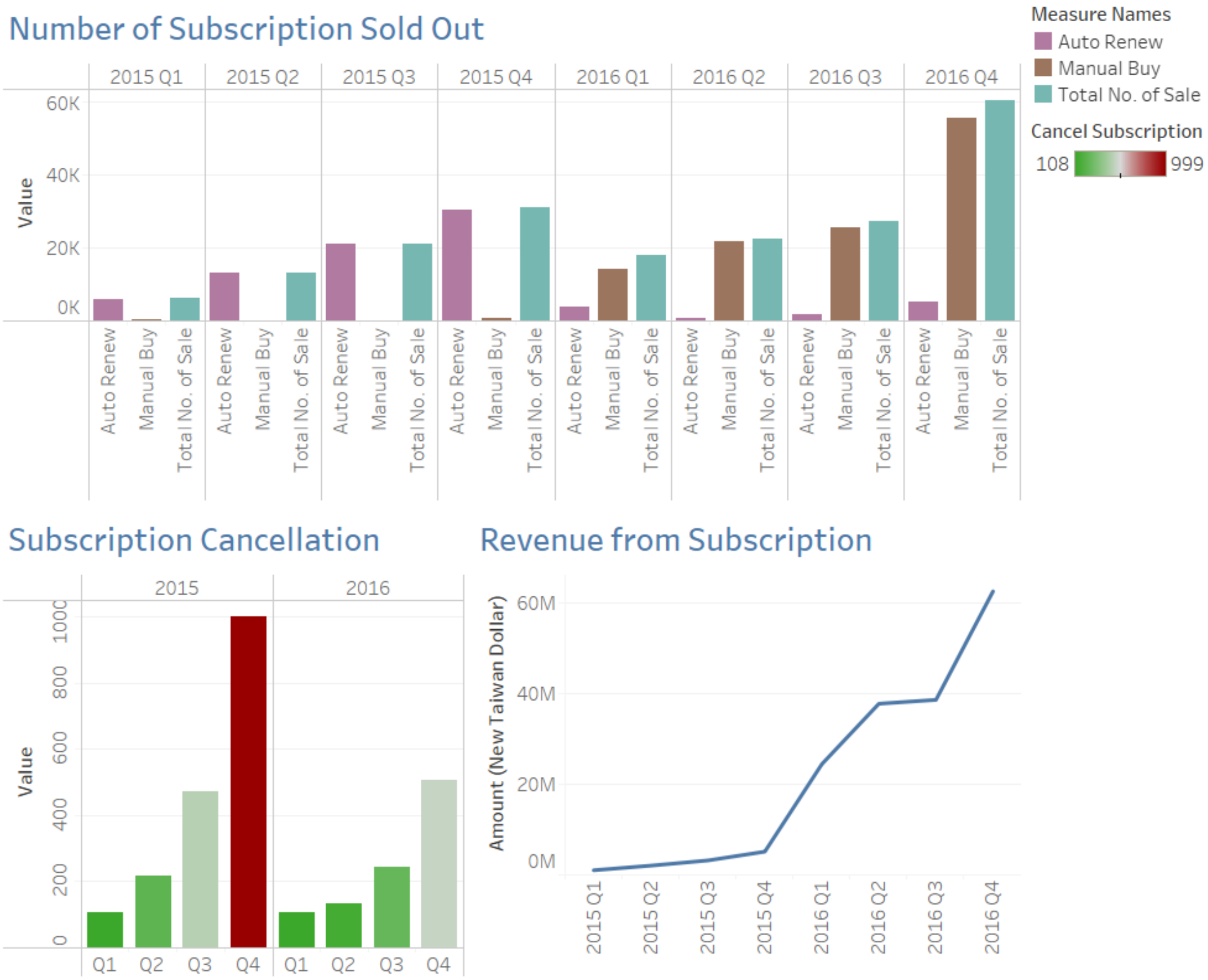
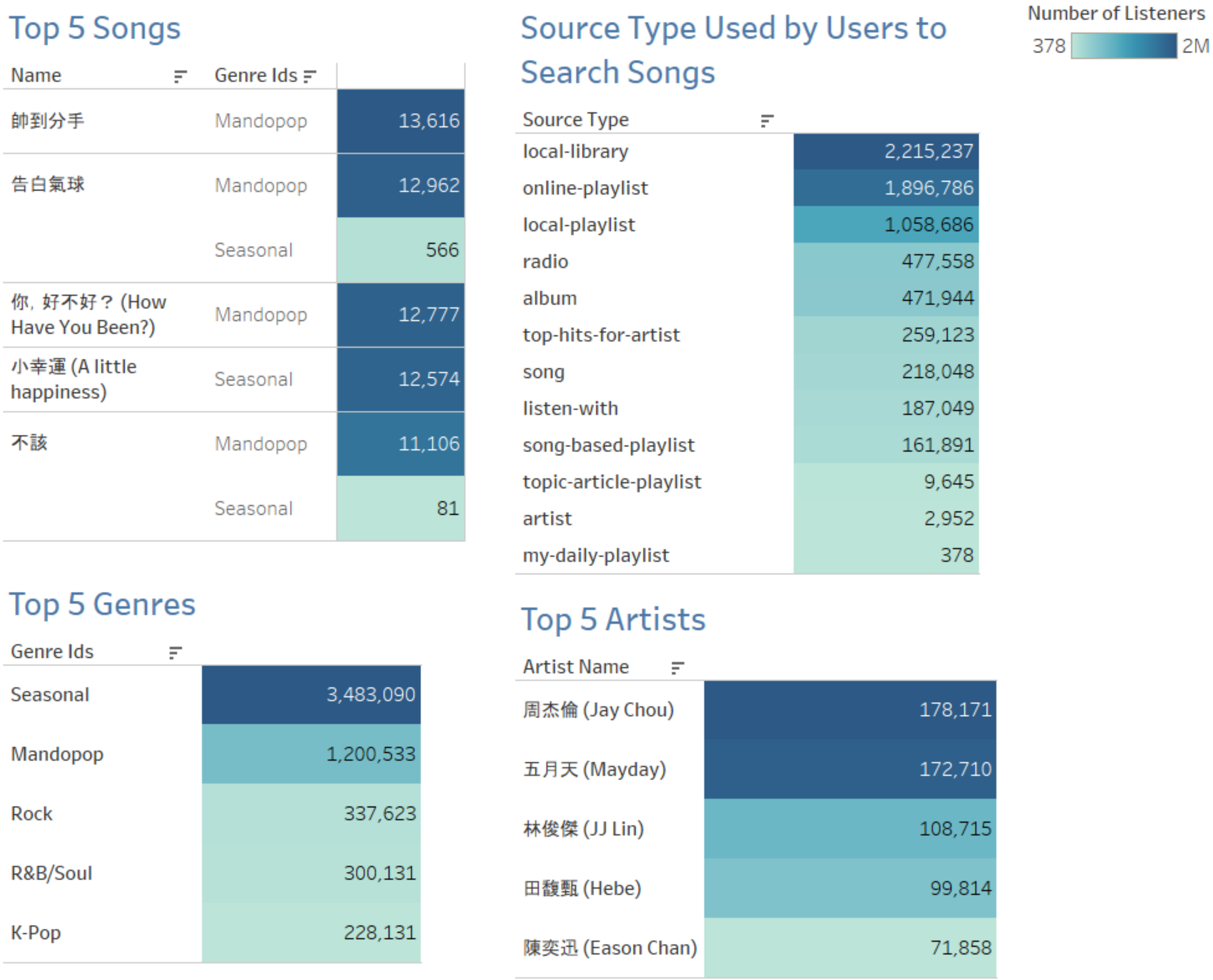


Figure 9 : Sale Performance Analysis Dashboard

This dashboard displays the Sales Analysis Dashboard, which is based on the number of subscriptions sold out, membership cancellations per year, and net profits per year. Auto renew users climbed from the first quarter of 2015 to a peak of 30K at the end of the year 2105, as seen in the top graph. However, from the beginning of 2016 to the end of the year, there was a significant reduction. Then, there were no manual buy users in 2015, but they gradually rose from the beginning of 2016 to a maximum point (55K) at the closing of the year. Consequently, from the beginning of 2015 to the end of 2016, there was a fluctuation in the overall number of sales.

We can observe from the subscription cancellation graph that the end of the year had more users canceling their subscriptions than the first quarter of the year. However, the membership plan was canceled for 999 users in the last quarter of 2015, which is the highest rate this year. As indicated in the trend line graph, revenue appears to have risen dramatically year after year, with a high revenue from subscriptions in 2016.

This dashboard can be used to quantify the profitability ratio. By disclosing this demographic information to the public, the music streaming service can provide discounts and adverts to its consumers.



Top 5 Songs

Name	Genre Ids	
帥到分手	Mandopop	13,616
告白氣球	Mandopop	12,962
	Seasonal	566
你，好不好？(How Have You Been?)	Mandopop	12,777
小幸運 (A little happiness)	Seasonal	12,574
不該	Mandopop	11,106
	Seasonal	81

Source Type Used by Users to Search Songs

Source Type	
local-library	2,215,237
online-playlist	1,896,786
local-playlist	1,058,686
radio	477,558
album	471,944
top-hits-for-artist	259,123
song	218,048
listen-with	187,049
song-based-playlist	161,891
topic-article-playlist	9,645
artist	2,952
my-daily-playlist	378

Top 5 Genres

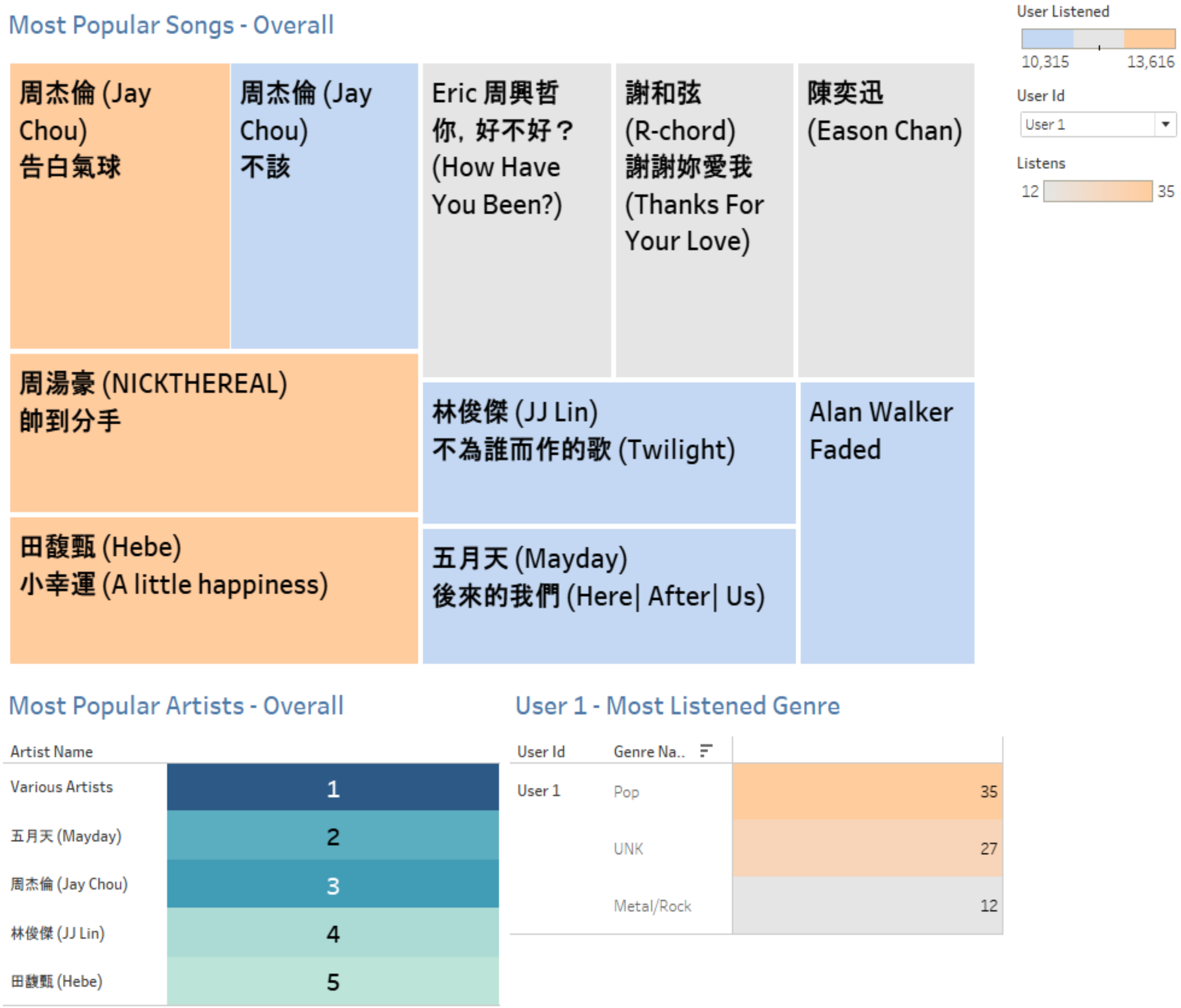
Genre Ids	
Seasonal	3,483,090
Mandopop	1,200,533
Rock	337,623
R&B/Soul	300,131
K-Pop	228,131

Top 5 Artists

Artist Name	
周杰倫 (Jay Chou)	178,171
五月天 (Mayday)	172,710
林俊傑 (JJ Lin)	108,715
田馥甄 (Hebe)	99,814
陳奕迅 (Eason Chan)	71,858

Figure 10 : Top songs, artists and genres Dashboard

The top5 songs, genres, and musicians are listed in the dashboard above, along with the most popular source type used by users to search for music. The most common choice song has roughly 13k users, but the second and third ranks only have around 12k. Seasonal and mandopop have the most users, while mandopop has the second highest users. K-pop is in the bottom place with 228, 131 people among the top five genres. Furthermore, among the top five artists, Jay Chou is the most popular, while Eason Chan is the least popular. Moreover, 2,215,237 users and 1,896,786 users used to search the music from the local library and online playlist, respectively. Their everyday playlist, on the other hand, has the fewest searches.


Most Popular Songs - Overall
周杰倫 (Jay Chou) 告白氣球
周杰倫 (Jay Chou) 不該
Eric 周興哲 你，好不好？(How Have You Been?)
謝和弦 (R-chord) 謝謝妳愛我 (Thanks For Your Love)
陳奕迅 (Eason Chan)
周湯豪 (NICKTHEREAL) 帥到分手
林俊傑 (JJ Lin) 不為誰而作的歌 (Twilight)
Alan Walker Faded
田馥甄 (Hebe) 小幸運 (A little happiness)
五月天 (Mayday) 後來的我們 (Here| After| Us)
User Listened
10,315
13,616
User Id
User 1
Listens
12
35

Most Popular Artists - Overall

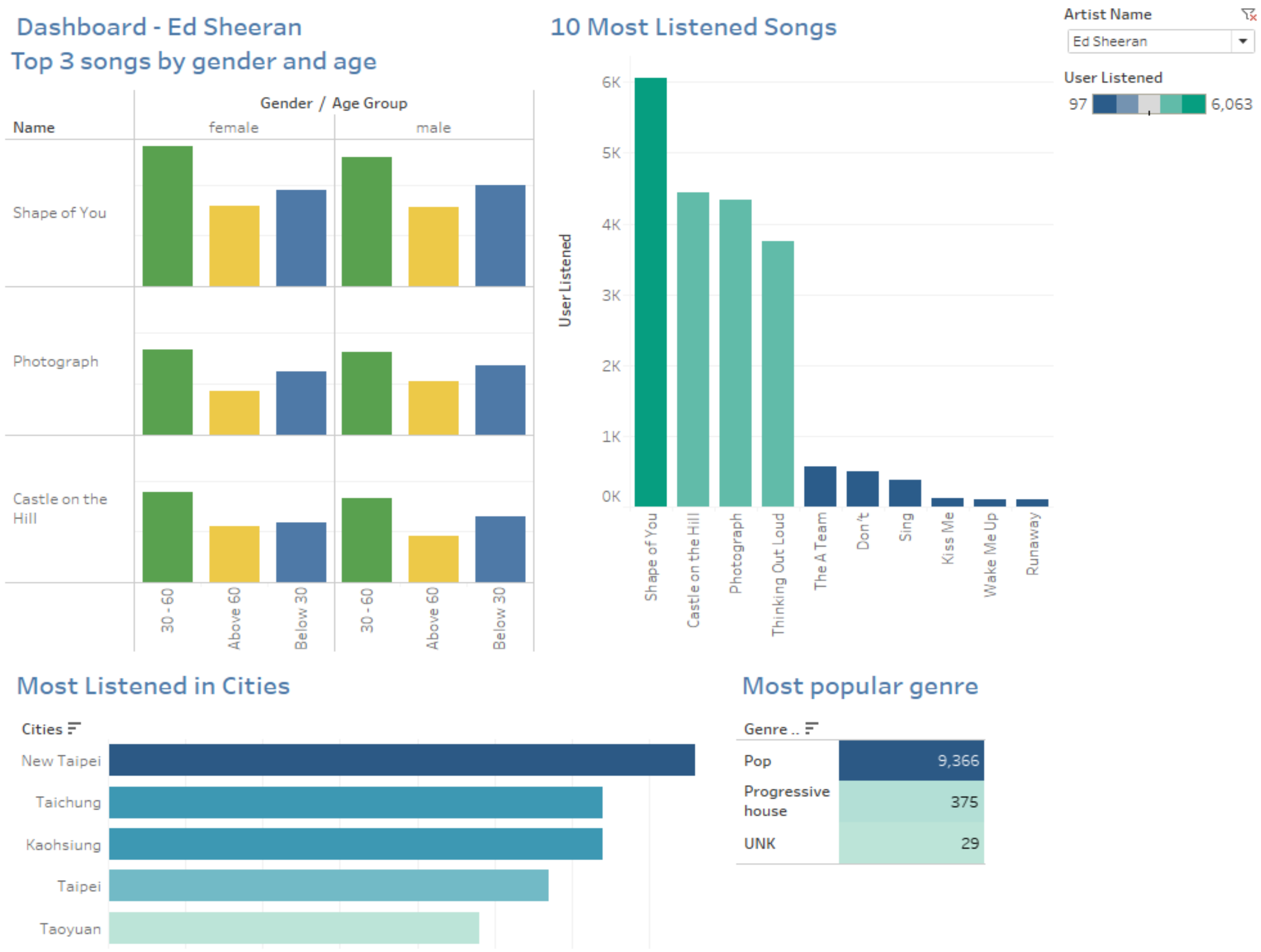
Artist Name	
Various Artists	1
五月天 (Mayday)	2
周杰倫 (Jay Chou)	3
林俊傑 (JJ Lin)	4
田馥甄 (Hebe)	5

User 1 - Most Listened Genre

User Id	Genre Na..	
User 1	Pop	35
	UNK	27
	Metal/Rock	12

Figure 11 : User/ Audience Dashboard

The platform provides the user with an overview of the current popular trends in the platform. It shows the top 10 most listened songs available in the platform and the most popular artists in the platform which can give the audience the general idea of what is trending and what trends to follow. It also gives the users information about which genre of songs they are listening to the most so that they can track their taste and activity on the platform.



Genre ..	
Pop	9,366
Progressive house	375
UNK	29

Figure 12 : Artist Dashboard

The platform provides the artists with a sophisticated dashboard. The dashboard contains the top 3 most listened songs of the particular artists according to their age and gender demographic, this allows the artist to know their audience better. The top 10 most listened songs of the artist shows which song is more loved and listened by the people and which songs are not so popular on the platform. With these two pieces of information, the artist can know which audience should be targeted to increase their success. In addition to that, the dashboard consists of which city that the artist is most listened to. This would help them know where most of their fanbase are located. It can help them make decisions on which city they should tour, organize events, concerts etc. The dashboard also shows what genre of songs that the artist has produced is more popular among the fans. This information can help them plan and calculate the risks on their future projects.

Predictive Analysis

Prediction Models and Analytics

For this project we created a simple classification model to predict whether a user will like a given song or not. This was measured by whether the user listened to the said song again or not within a month from the song user relationship which was labeled as target in the dataset. To train the model we combined the 3 tables, Song_info, user_info and song_user_relation by using the song ID and user ID of the songs and users respectively. All categorical data was converted to numerical using the pandas library. Then 7 features were extracted which seemed to affect the target value the most. These were

1. Song length
2. Song Genre
3. Song language
4. Source type (where the listener heard the song)
5. User Age
6. User City
7. User gender

Various classification machine learning methods were used with a split of 70 train and 30 test and XGBoost was found to have the highest test accuracy of 64% when predicting the target using the above 7 features.

Model	Test Accuracy (percent)
Logistic Regression	60
GaussianNB	57
XGBoost	**64**
Decision tree	59

Tabel 1: Comparison of ML Models

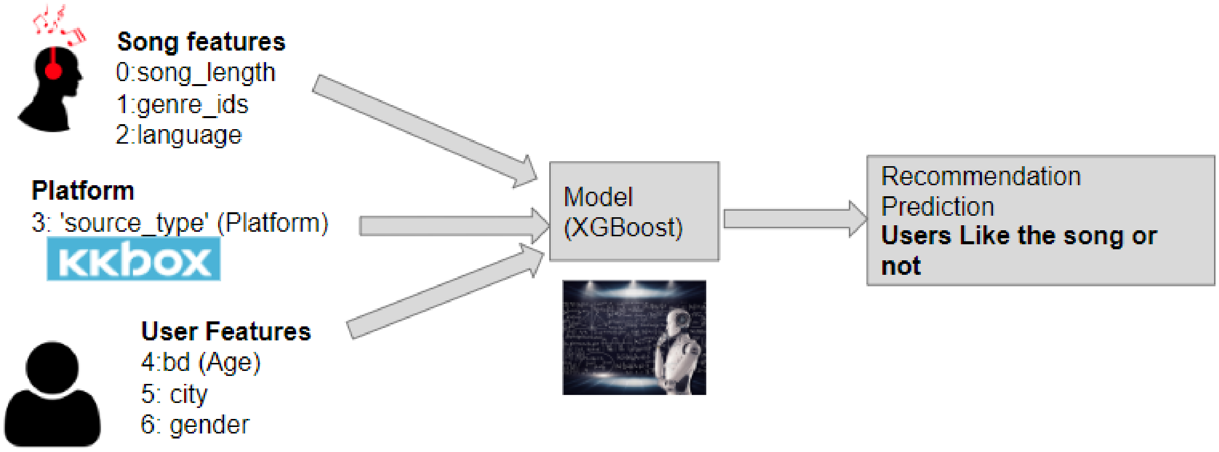
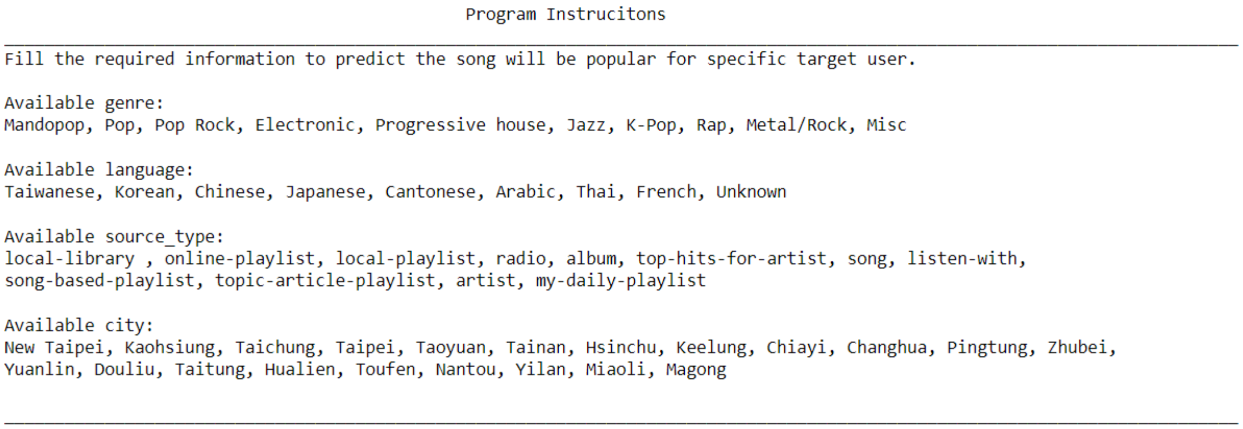


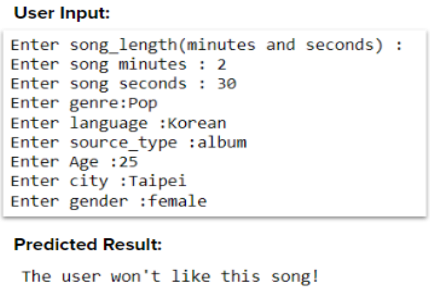
Figure 13 : The prototype of prediction system implementation

So the way the prediction model would work would be: when we input the above 7 features into the model with the options as shown in the Program Instructions, it will output whether the said user will like the given song or not.



```
                              Program Instrucitons
________________________________________________________________________________________
Fill the required information to predict the song will be popular for specific target user.

Available genre:
Mandopop, Pop, Pop Rock, Electronic, Progressive house, Jazz, K-Pop, Rap, Metal/Rock, Misc

Available language:
Taiwanese, Korean, Chinese, Japanese, Cantonese, Arabic, Thai, French, Unknown

Available source_type:
local-library , online-playlist, local-playlist, radio, album, top-hits-for-artist, song, listen-with,
song-based-playlist, topic-article-playlist, artist, my-daily-playlist

Available city:
New Taipei, Kaohsiung, Taichung, Taipei, Taoyuan, Tainan, Hsinchu, Keelung, Chiayi, Changhua, Pingtung, Zhubei,
Yuanlin, Douliu, Taitung, Hualien, Toufen, Nantou, Yilan, Miaoli, Magong

________________________________________________________________________________________
```

Figure 14: Program instruction for prediction system implementation



User Input:

```
Enter song_length(minutes and seconds) :
Enter song minutes : 2
Enter song seconds : 30
Enter genre:Pop
Enter language :Korean
Enter source_type :album
Enter Age :25
Enter city :Taipei
Enter gender :female
```

Predicted Result:

```
The user won't like this song!
```

Figure 15 : Prediction result

The main usage of this model is to be able to predict whether the said user will like a song or not. Then recommend a song that the user might like (according to the model) to that user. For a user the idea is to scan the entire song catalog to find the best matches so that the recommended song is the one the user will listen to again.

The features we have used in this project are limited as there is not much quantitative data about the song such as *loudness*, *beats per minute*, *time signature* etc. but still with the current features a certain level of prediction and its concept can be understood.

Prescriptive Analysis

From the above descriptive and predictive analysis we will be able to recommend the three end users with the ability to make decisions that could help improve the engagement of the overall platform with the audience. Each end user plays a different role in the platform hence each of them get a different dashboard, hence a different recommendation.

Recommendation to Company

The above dashboards help show the highest user demographic which can be leveraged to target ad companies which are interested in a specific demographic. The highest demographic could be charged the most accordingly. For example: If a company wants to show an ad targeted at young audiences and our platform has a large young audience then it would be lucrative for that company to use our platform to show their ads.

We can also analyze when subscriptions are being canceled the most and find out the reason why to try and mitigate this from happening. Similarly the company could promote most attractive subscriptions plans in its own advertisements based on the subscription data.

We could analyze the top trending music on the platform and then recommend these trending music more for higher audience engagement on the music platform. Also finding out about the most popular source type for the audience would help put trending songs in the most popular source type for higher audience engagement with the most liked songs.

Recommendation to Artist

For the artist the above insights could help them know what is their target audience. The gender and age group of their target audience could the artists write lyrics and songs with more relevance to this demographic. Also, we give the top trending song and the genre of the artist so they can know what type of songs perform on this platform. We also show the top cities the artist is famous in so they can plan their tours and live sessions accordingly.

Recommendation to Users

Users are the main end users of the platform so the main aim is to get them more engaged. To the users we show the top trending songs/ albums/ artists in the platform so they can listen to what is being liked by most people on the platform. We also show them their current listening streak and number of listens so they can know their own tastes. Finally we use our predictive ‘song likeness’ model and recommend songs to the user which they might like the most, hence increasing user engagement.

Conclusion

Music streaming platforms aim for maximum customer engagement by increasing the duration of listening. This is what our BIS is aiming for. This could be increased by analyzing the data and providing various insights through visualizations and predictions which eventually help

1. Recommending the right songs to the user,
2. Showing which songs are performing well in the platform to the artists so they can continue creating those type of songs
3. Showing overall high performance songs in the platform in order to promote these artists and songs to more listeners.

Future Works

We will create a website with the best user interface for the convenience of our users. In this website, we will have three user sections for company, artist and listeners. For the company section, we will show the dashboards such as the user demographic dashboards, sales performance analysis dashboard and operational analysis dashboard, and our analysis result. And the predictions of the type of songs which will be liked by a specific listener type. We will clearly describe the inputs data for the ease of use for them. There will also be a feature where the company can add the data from other countries, so they can predict which types of songs will be popular for specific types of listener groups in those countries.

For the artist section, we will show them their top song and genre, and the visualization of their top songs listened by each age group and gender. And we will provide them with the recommendations to create which music genres to attract more users. For the listener section, we will show the information about the top songs and artists overall. In addition, we will personalize the data by using their gender, age, locations, their most listened songs, genres and artists. then give recommendations which songs they should listen to.

References

1. The dataset sources 1: WSDM - KKBox's Music Recommendation Challenge: https://www.kaggle.com/c/kkbox-music-recommendation-challenge
2. The dataset sources 2: WSDM - KKBox's Churn Prediction Challenge https://www.kaggle.com/c/kkbox-churn-prediction-challenge/data?select=members_v3.csv.7z
3. Medium article on WSDM — KKBox’s Music Recommendation Challenge
4. KKBox’s Music Recommendation by Yunru Huang, Mengyu Li and Yun Wu